



# AI 응용시스템의 이해와 구축

3강. 데이터 콜렉션, Label Quality, Responsible AI

---

# 데이터 수집 (Data Collection)

# Define Modeling Data (2강)

예) 쇼핑웹사이트의 상품 추천

모델링을 위한 데이터  
(User Data)

- 상품 구매 이력
- 사용자의 **demographic** 정보
- 사용자의 **geographic** 정보

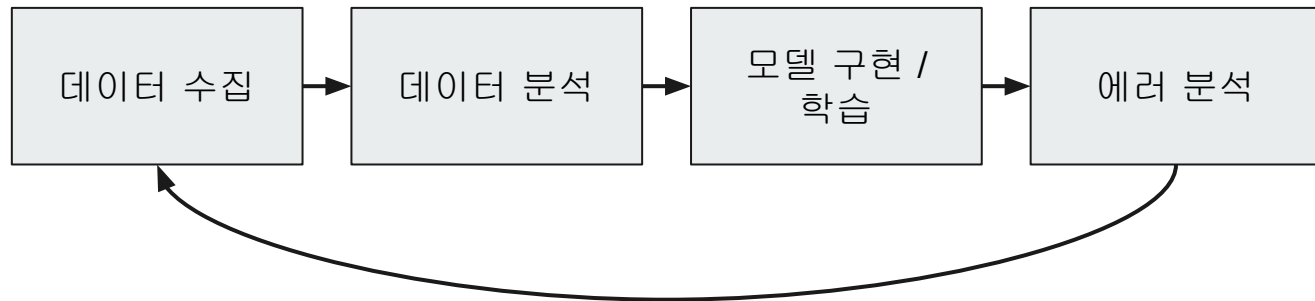
Feature Data

- 사용자 **demographic**: age, gender, racial\_info
- 거주지 정보: country, state, zip code
- 검색 이력: (search query, search date)
- 상품 추천 이력: (product ID, recommended date)
- 상품구매 이력: (product ID, purchase date)

Label Data

- 상품 구매 이력: (product ID, purchase date)
- 구매 의사 이력: (click through, date)
- 비호감 의사: user clicked, but exited
- 검색에서 구매까지의 # of clicks

# Data Collection Process



- **데이터 수집**: 효과적인 Feature + Label을 얻기까지 여러번의 데이터 수집이 필요.
- 데이터 수집에 (ex. “완벽한 데이터”를 얻기 위해) 오랜 시간을 투자한 후 모델 구현을 하기보다는 최소한의 데이터를 먼저 수집 후 **baseline** 모델을 사용하여 데이터 교정을 추천.
- 예) 데이터 수집에 3주, 모델구현에 1주, 에러분석에 1주  
수집 1주, 구현 1주, 분석 1주, 반복

X

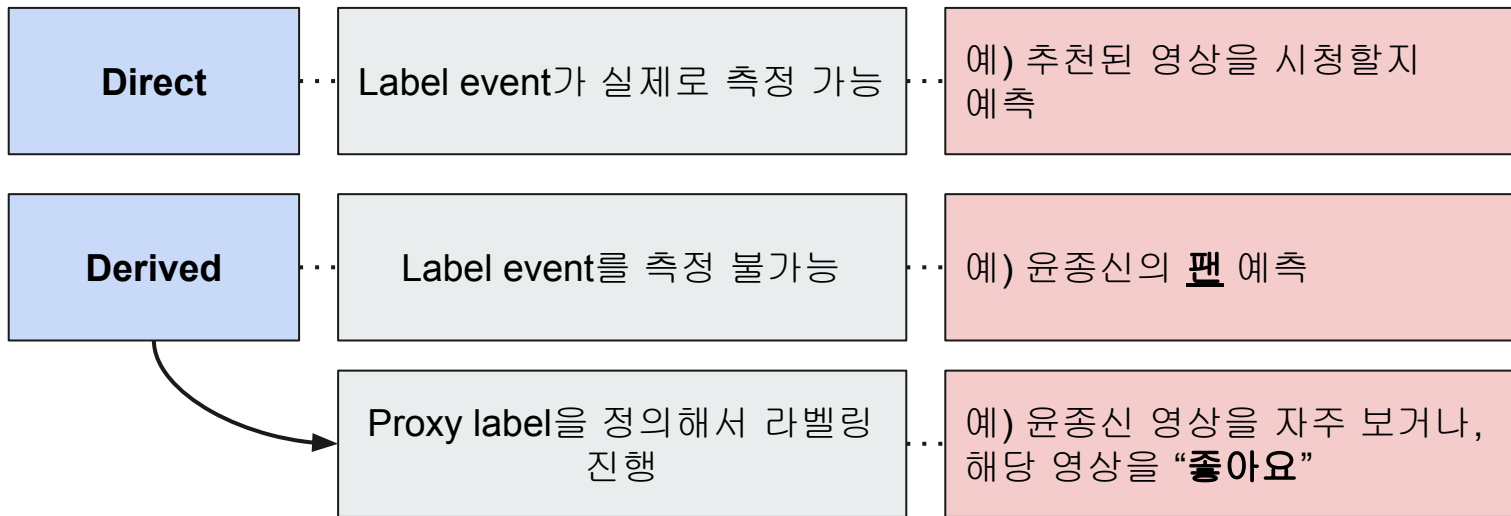
O

# 라벨(Label)의 종류들

## Direct Label vs. Derived Label

예측하려고 하는 변수를 측정가능(Observable) 한가?

예) 유튜브에서 ML모델들



“윤종신의 팬”  
데이터는 구하기  
매우 어려움.

# Data Labeling Methods

---

어떤 타입의 라벨링 방법이 있나?

- **피드백 라벨링 (Feedback Labeling)**
  - 시스템에 생성되는 유저피드백으로 라벨링
  - 종종 Derived Label을 이용하여 진행.
- **휴먼 라벨링 (Human Labeling)**
- Semi-supervised labeling
- Active learning
- Weak Supervision

이 외에도 다양한  
라벨링 방법들 존재.



# 피드백 라벨링의 장단점



## 장점

- 연속적으로 새로운 학습데이터가 생성된다.
- **Direct label**일 경우 라벨이 유기적으로 **evolve**된다.
- 파워풀한 시그널 (실제 사용자의 클릭!)

## 단점

- **Derived label**일 경우 “Ground Truth”에서 점점 멀어질 수 있음.
- 이미 **deploy**된 시스템 필요
- **Log data**를 **training data**로 변환하는 시스템 필요.

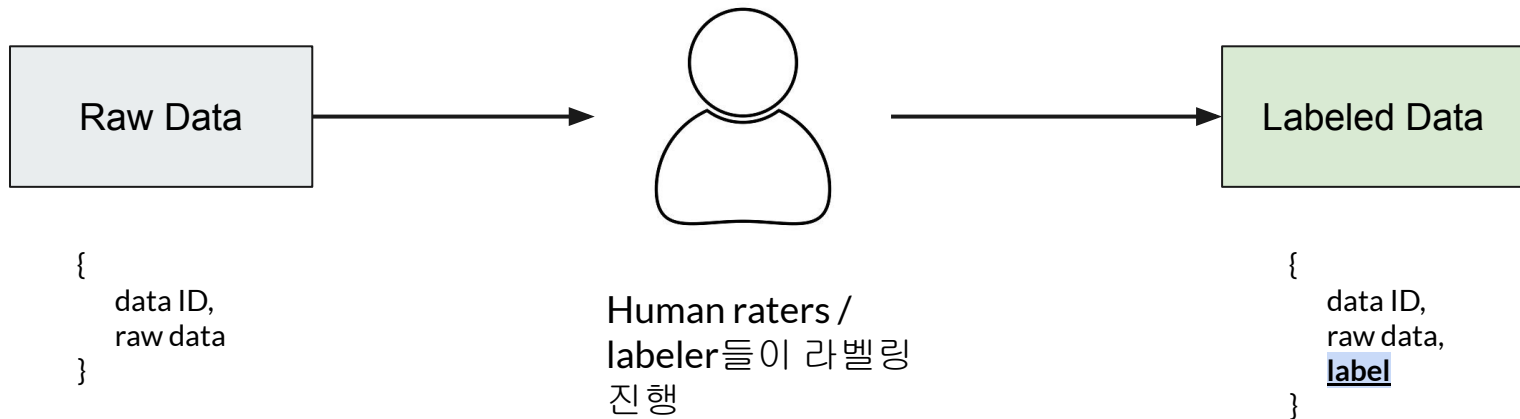


# Human Labeling

구글에서는  
**Rater**라고  
불리기도 함.  
[search quality  
rating](#)

라벨러(Labeler): Human Labeling을 진행할 때, 데이터 라벨 (label)을 생성하는 사람을 지칭.

매우 매뉴얼한 라벨링 작업.



# 휴먼 라벨링의 장단점

## 장점

- Derived (proxy) label일 필요 없이 제품이 필요로하는 라벨을 수집할 수 있다!
  - 사진에서 “강아지”, “고양이” 라벨.
- 베이스라인 시스템이 없어도 **Boostrapping!**

Feedback Labeling에 비해 큰 데이터 수집은 어려움.

## 단점

- Labeling task가 비쌀 수 있음.
  - 엑스레이 판독.
- 데이터가 클 때는 (high dimensional data) 라벨링 불가능.
  - 오랫동안 사용자 히스토리를 보고 선호도 추정.
- 품질이 불규칙 할 수도.
- 데이터 수집의 느린 속도.

# 라벨 수집은 누가 할 것인가?



일반적으로 세가지 옵션이 존재:

1. 인하우스: 개발자(ML엔지니어)가 직접
2. 클라우드소스: 도메인 전문가 (domain expert, subject matter expert)에게 위탁
3. 아웃소스: 외부 데이터 취득

## 라벨 수집은 누가 할것인가?

	장점	단점
인하우스	라벨 정의에 높은 이해	높은 <b>cost</b> (엔지니어) 때로는 도메인지식 부족
클라우드소스	낮은 <b>cost</b> (일반적인 데이터 문제)	도메인지식 부족하거나 그렇지 않은경우 높은 <b>cost</b>
아웃소스	오픈소스 데이터 때로 낮은 <b>cost</b>	다양하지 않은 <b>label</b> 정의

## 라벨 수집은 누가 할것인가?

	사진속에 사물 감지 (Object Detection)	의료 진단 노트에서 질병 진단
인하우스	가능. 비쌈. ❌	도메인지식 부족. ❌
클라우드소스	✅	도메인지식 가진 라벨러 구하기 어려움 🤔
아웃소스	✅	✅

## 라벨 수집은 누가 할 것인가?

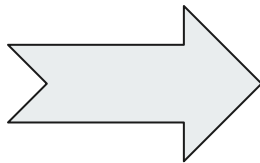
	사진속에 사물 감지 (Object Detection)	의료 진단 노트에서 질병 진단
인하우스	가능. 비쌈. ✗	도메인지식 부족. ✗
클라우드소스	(플랫폼) <a href="#">Google Crowdsourcing</a> <a href="#">Amazon Mechanical Turk</a> (솔루션) <a href="#">Bobidi</a> , <a href="#">LabelBox</a> , <a href="#">Datamaker</a>	의료인력
아웃소스	(오픈소스) CIFAR-10, ImageNet,... LISA Traffic, DeepFashion,...	<a href="#">i2b2</a> <a href="#">icd9</a> , icd10

# 휴먼 라벨링의 예: 의료 진단서로 질병 판독

Problem: Medical Note(의료 진단서)를 읽고 환자의 질병 및 진단을 정형화하는 문제.

“ptn with hbp, with dm2 risks, 60m”

“60yr male, has symptoms of htn, with potential type2 diabetes”

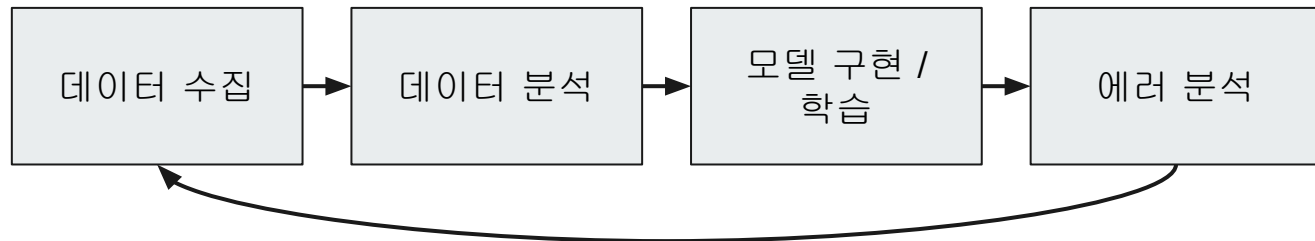


Age: 60  
Gender: Male

Condition: Type 2 diabetes  
Severity: LOW

Condition: Hypertension  
Severity: HIGH

# Data Collection Process



- 시작은 가장 **cost effective**한 데이터셋으로 데이터셋을 구성.
  - 예) 아웃소싱한 오픈소스 데이터로 시작.
  - 아웃소싱 데이터가 현 문제의 정의와 다르다면 레이블 수정 필요.
- 데이터 분석 및 모델 에러 분석을 통해 부족한 데이터 확인
  - 새로운 데이터 소스를 취득해 기존 데이터와 merge.
  - Merge할 때 레이블 분석 + 모델 분석 필요.

새로운 데이터를 수집 시 기존 데이터 대비 너무 큰 데이터셋을 추가하지 않도록 권장.

- 데이터셋이 10배 이상 커지면 커다란 모델 + 데이터 분석 차이를 초래.
- 늘 **iterative**하게 조금씩 데이터 확장을 추천.



# 환자 질병 판독 문제

“60m ptn with hx htn and dm2 symptoms”

- 우선 outsource (ICD9)을 통해 데이터 수집.

ICD9Data.com

Search

Home > 2012 ICD-9-CM Diagnosis Codes > Diseases Of The

**Hypertensive Disease 401-405 >**

- 401 Essential hypertension
- 402 Hypertensive heart disease
- 403 Hypertensive chronic kidney disease
- 404 Hypertensive heart and chronic kidney disease
- 405 Secondary hypertension

ICD9Data.com

Search

Home > 2012 ICD-9-CM Diagnosis Codes > Diseases Of The Circulatory Sys

**Essential hypertension 401- >**

- Hypertension occurring without preexisting renal disease or known on

▶ 401 Essential hypertension

▶ 401.0 Malignant essential hypertension [convert 401.0 to ICD-10-CM](#)

▶ 401.1 Benign essential hypertension [convert 401.1 to ICD-10-CM](#)

▶ 401.9 Unspecified essential hypertension [convert 401.9 to ICD-10-CM](#)

# 환자 질병 판독 문제

“60yr old male has history of hypertension and diabetic symptoms”

Data Source	
ICD9	401 <b>Essential hypertension</b> 402 <b>Hypertensive heart disease</b> 403 <b>Hypertensive chronic kidney disease</b> 405 <b>Secondary hypertension</b> ....

“60m ptn with hx htn and dm2 symptoms”

htn, hbp, high blood pressure 같은  
동의어 부족!

# 데이터 수집의 예

추가로 크라우드 소싱 진행.

실제 의료인들에게  
crowd sourcing 의뢰

Task: 다음 문장에 등장하는 질병들의  
토큰과, 그에 해당하는 ICD9 코드를  
고르시오.

“60m ptn with hx **htn** and **dm2**  
symptoms”

“htn”, 401 **Essential hypertension**,  
“dm2”, E11 **Type 2 diabetes melitus**

# 데이터 수집의 예

“60yr old male has history of hypertension and diabetic symptoms”

Data Source	
ICD9	401 <b>Essential hypertension</b> 402 <b>Hypertensive heart disease</b> 403 <b>Hypertensive chronic kidney disease</b> 405 <b>Secondary hypertension</b> ....
Crowdsource	htn, hbp, high blood pressure, ...

# Data Collection Summary



- 라벨러 (Labeler)들의 expertise를 고려.
- 라벨러들의 다양성을 고려 (시스템 유저들과 가능하면 비슷한 분포)
- 라벨링의 Cost를 고려
- 라벨링된 데이터가 Fresh해야 하나?
  - 그렇다면 주기적으로 반복하여 라벨링 진행.

---

# Quiz & Final Project Team Assignment

## Quiz 3



Quiz 3 링크

# Project Teams: Cluster + Pitch New Project

프로젝트	팀원
의료진단	조우성, 안혜영, +1
상품 추천	노용문, 신용주, 박준호, 한진웅,
웹 검색	
스마트 스피커 음성인식	권용순
제조공정 불량 판정	임향빈, 오동규, 권용순, 신용주, 박주형, 안혜영, 정근시
제품 리뷰 (긍정/부정)판독	박주형, 정운국, 김용욱
영상추천	
광고 추천	
다른 프로젝트...??	

- 3/20일 자정까지 학생분들끼리 조정하여 슬랙에 업데이트.
- 차후 강사가 팀 조정 예정.



---


# Data: Label Quality

# Data Quotes



Data is the hardest part of ML, and most important piece to get it right.

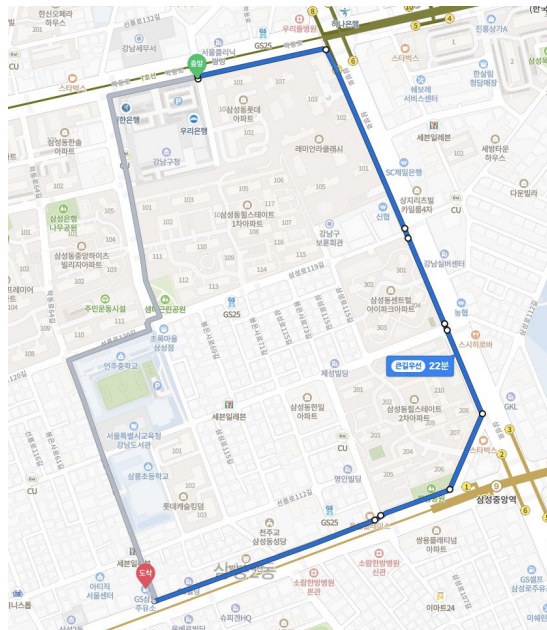
*(Michaelangelo, Uber)*



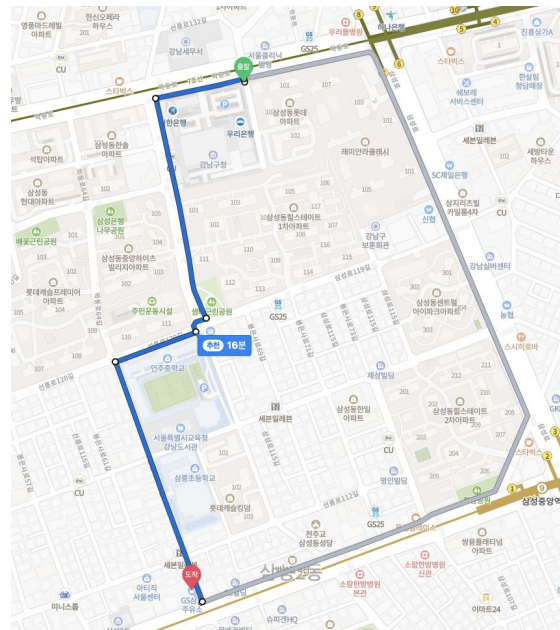
근데 왜?

# Human Evaluation

예) 장소 A에서 B까지 추천하는 길은?



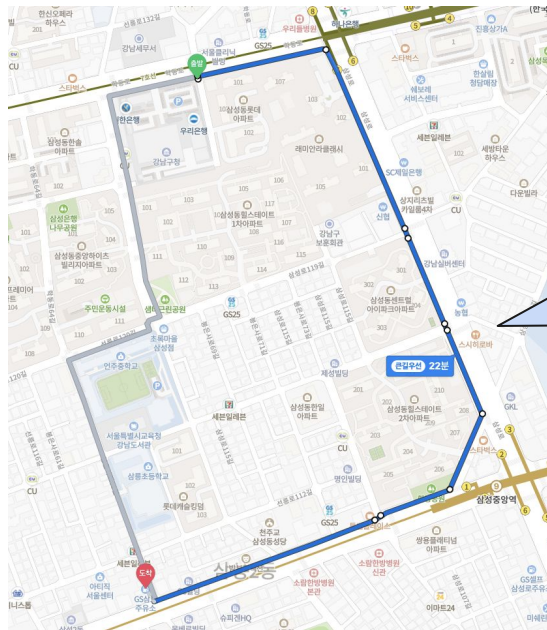
Ground Truth



사람 X

# Human Evaluation

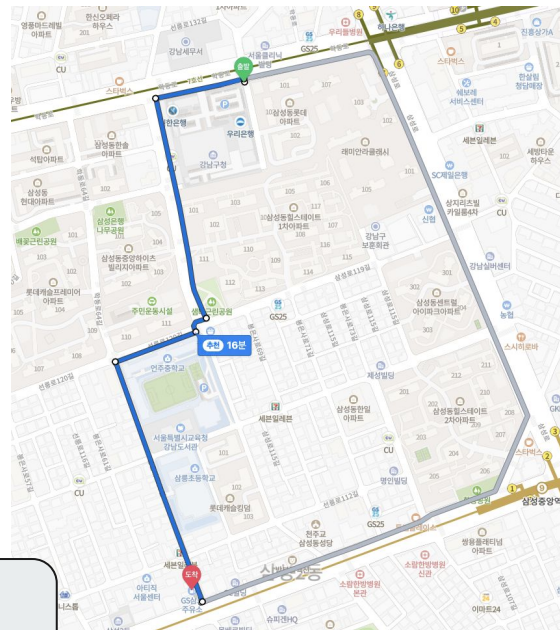
예) 장소 A에서 B까지 추천하는 길은?



Ground Truth

큰길  
선호

Ground  
Truth도 사람  
Y가 만든  
데이터!

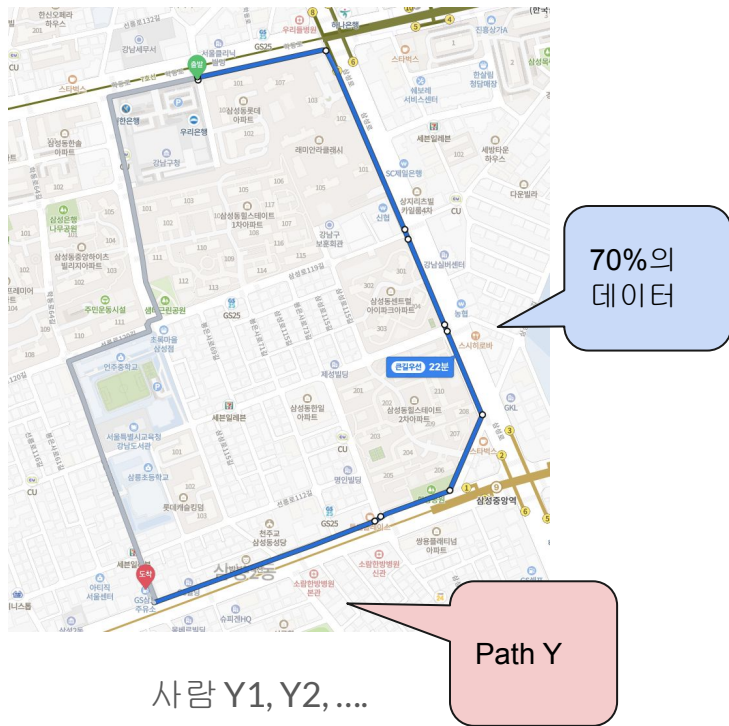


사람 X

도보용  
골목길  
선호

# Human Evaluation

예) 장소 A에서 B까지 추천하는 길은?



# Human Evaluation

예) 장소 A에서 B까지 추천하는 길은?

2명의 사람에게 이 질문을 했을때 동일한 길을 말하는 확률은?

$$0.7^2 + 0.3^2 = 0.58$$

이러한 **ground truth**로 구성된 학습데이터로 모델을 만들면 모델의 정확도는?

0.7

Naive Bayes  
Classifier

우리 모델이 인간보다 더  
정확한가?  
No!

# How to deal with human errors?

그냥 더 많은 데이터 수집을 하면  
되나?

No. 그냥 다른 분포의 데이터만  
초래.

Ground truth가 정말 error free한다면?  
예) click prediction을 위한 클릭 데이터

Human prediction이 **Bayes**  
**Error**를 추산

# 베이지 에러 (Bayes Error)

“Irreducible Error”

Classification 문제에서, Training data를 완벽히 학습하였을 때 (즉, 그 사건의 underlying density function 을 알고 있을 때), 그 Training data에 대해 가장 확률이 높은 Class Label을 선택하는 방법에서 발생하는 이론적 최소 오차.

$$1 - E \left( \max_j \Pr(Y = j | X) \right)$$

$P(y | x) = x$  라는 인풋을 받고  $y$ 라는 결과가 나올 확률.

예) “A에서 B까지 추천하는 길”의 종류가  $n$ 개가 있을 때, 가장 흔히 선택하는 길을 제외한 나머지 길들을 선택할 확률.

“A에서 B까지 추천하는 길”의 종류가 적거나, 한 길이 명확할수록 낮은 Bayes Error

ML문제에서 동일한  $X$ 를 주었는데 다른 결과가 나오는 이유?

$X$ 가 인풋을 완벽히 표현하지 못하기 때문..



# Model Error Type

**Input Feature X**를 사용하여 **Y**값을 예측하는 모델링을 할 때, 대부분의 경우 **X**가 **Y**를 완전히 표현할 수 없다.

따라서 **X**와 독립적인 변수  $\epsilon$ 가 **Y**값에 영향을 미친다. 이를 반영하면:

$$Y = f(X) + \epsilon$$

모델링을 통하여  $f()$  라는 함수를 찾게되는데, 우리가 찾게되는 함수는  $f()$  에 근사한 함수를 찾는다.

$$\hat{Y} = \hat{f}(X)$$

여기서 두가지 에러가 발생:

ML 모델이 만드는  
에러

Reducible Error:

$$f(x) - \hat{f}(X)$$

Irreducible Error (Bayes Error):

$$\epsilon$$

Feature X가  
부족해서 생기는  
에러

# How to deal with human errors?

그냥 더 많은 데이터 수집을 하면 되나?

No. 그냥 다른 분포의 데이터만 초래.

Ground truth가 정말 **error free**한다면?  
예) click prediction을 위한 클릭 데이터

Human prediction이 **Bayes Error**를 추산

Ground truth가 human label인데 다른 인간과 **disagreement**가 보인다면?

라벨 수행법(Label instruction)을 좀 더 정교하게 재정의.

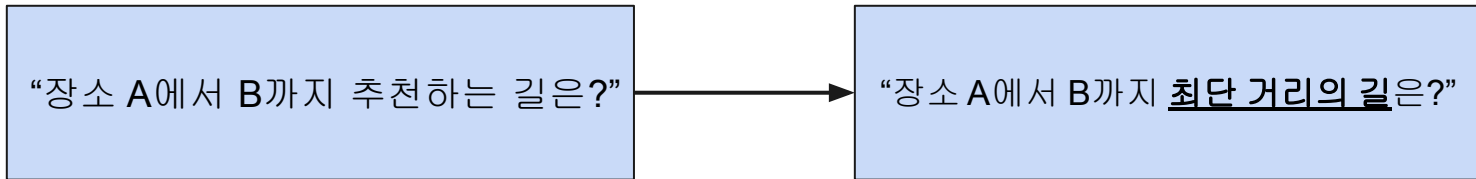
## Bayes Error:

$P(y | X)$ 의 확률 분포를 알고 있을 시, 이론적으로 도달할 수 있는 최소의 오차값.

e.g. 앞75%, 뒤25% 확률의 동전을 던질때, 늘 “앞면”이라고 예측해도 뒷면이 나올 확률

“장소 A에서 B까지 추천하는 길은?” → “장소 A에서 B까지 최단 거리의 길은?”

## 라벨 수행법 (Label Instruction) 재정의



이제 사람들끼리는 동일한 길 선택을 하게 됨.  
(Good consistent data!)

상대적으로 ML모델에게는 사람을 이기기  
어려워졌지만,  
좀 더 정교한 modeling problem!

# Label Quality Summary



데이터 라벨의 퀄리티가 좋지 않을 때엔, 문제가 쉽다는 오해를 불러일으킬 수 있다.

따라서 데이터 라벨의 퀄리티를 높이기 위해

- Labeling instruction을 수정하여 일관적인 라벨을 취득.
- 외부 데이터를 통해 ground truth 취득.

이에 따라 modeling problem이 어려워질 수 있으나, 모델링 알고리즘이 좀 더 공정한 metric data로 정확도를 판단할 수 있다.

# Feature: Ambiguous Input Data



데이터셋을 라벨러가 직접 보고도 라벨이 명확하지 않는 경우

예) 어시스턴트에게 명령

- [유리가 누구야?] (쿨, 핑클, 소녀시대, IZ\*ONE)
- [how old is washington] (도시, 주, 대통령)


# Feature: Ambiguous Input Data

때로는 labeler의 맥락에 따라 동일 데이터셋이 다른 의미.

- 예) 어시스턴트/검색엔진에 [Call Mom]

Task: Call


Modifier: Mom

 <https://www.yellowpages.ca> > bus

### Call Mom Designated Driver Service - 2340 Belair Rd, Victoria, BC


Call Mom Designated Driver Service - Victoria - phone number, website & address - BC - Taxis.

5.0 ★★★★★ (1)

 <https://www.canpages.ca> > page > ca...

### Call Mom Designated Driver Service - Victoria, BC - 2340 Belair Rd | Canpages

Please call 250-507-6515 to do business with Call Mom Designated Driver Service that is near your area. Finally, you can share this info with your contacts by ...


 <https://call-mom-drivers507-6515victoria-bc.weebly.com>

### Call Mom Designated Driver Service (250) 507-6515 - Home

[전체](#) [지도](#) [이미지](#) [도서](#) [동영상](#) [더보기](#) [도구](#)

검색결과 약 2,660,000,000개 (0.69초)

도움말: 한국어 검색결과만 검색합니다. 환경설정에서 검색 언어를 지정할 수 있습니다.

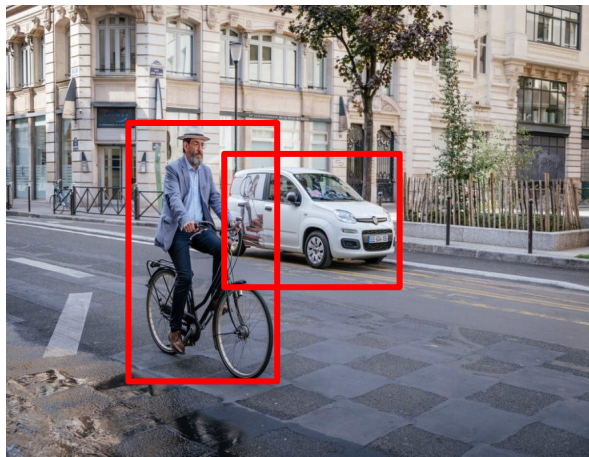


Olivia O'Brien - Call Mom (Audio) - YouTube

<https://www.youtube.com/watch>

# Label: Poor Label Format

- Object detection problem: 사진에서 “교통수단”이라는 물체를 감지하는 문제



- 어느 라벨이 옳은지는 product 응용에 따라 정함.
- 그러나 일관적인 포맷이 필요 (**Label consistency**)

# Data definition



인풋 데이터:

- 어떤 데이터를 feature로 사용할 것인가?

아웃풋 데이터:

- 어떻게 label을 consistent하게 유지할 것인가?

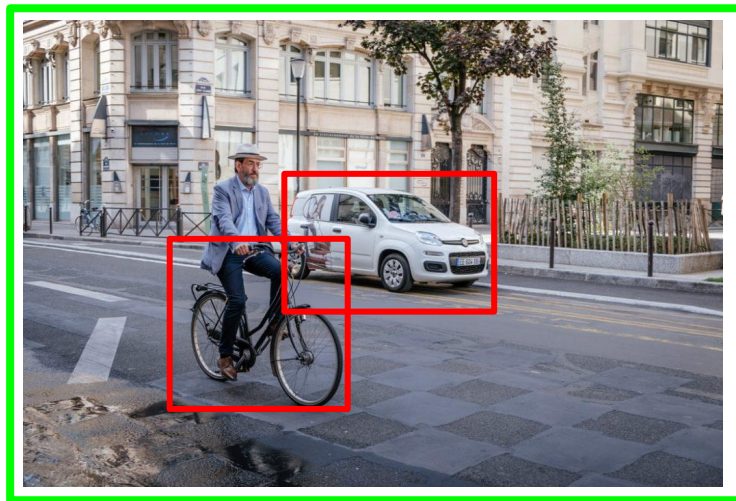
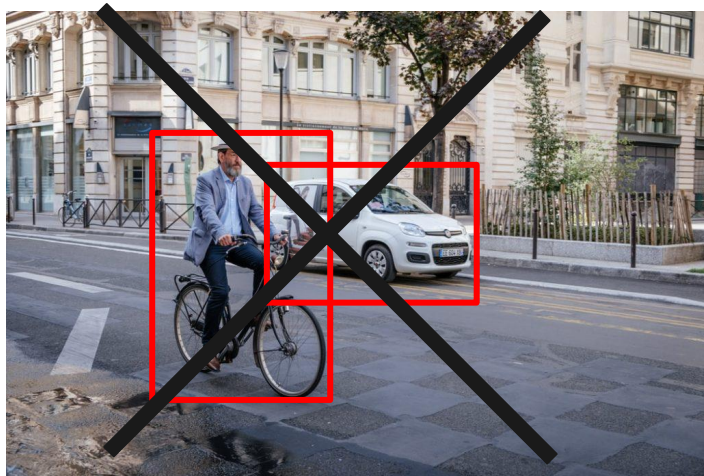


# Improve Label Quality

---

- 라벨러의 퀄리티:
  - Subject Matter Expert: 해당 데이터를 라벨링 할수 있는 사람들인가?
- 각 데이터 마다 라벨러를 여러명에게 라벨링 진행
- 라벨러 간에 전부 동의를 얻지 못하면?
  - 다수결로 결정.
  - Or, 모두 동의를 얻지 못한 데이터를 또 다른 라벨러에게 다시 라벨링 진행
  - Or, 라벨러들이 인풋에 정보가 부족하다고 판단하면
    - 인풋 데이터(feature)를 더 추가
- disagreed label을 줄일 수 있을때까지 계속 반복.

## 예: 라벨 정의의 표준화

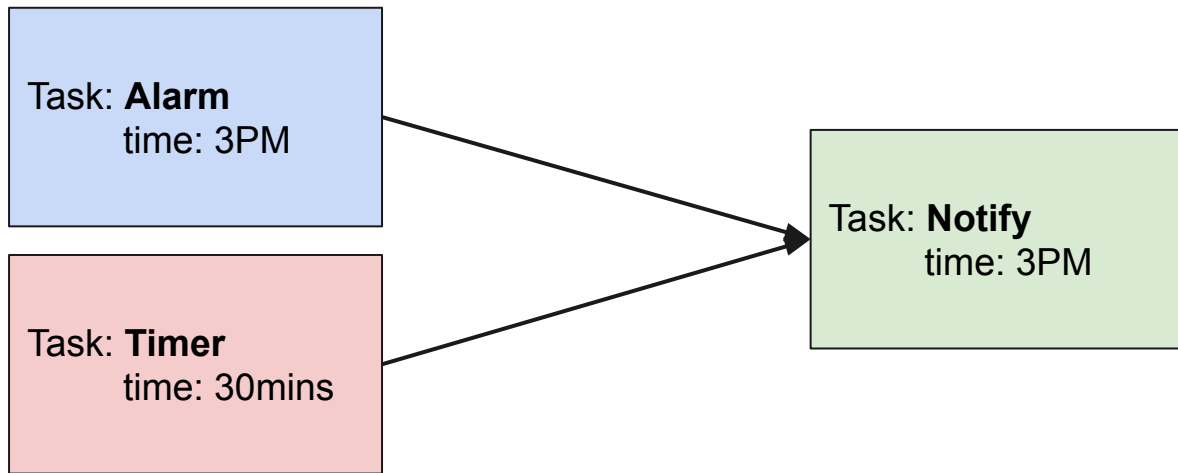


예) 서로 overlapping하더라도 명확히 vehicle의 bounding box를 라벨로 정하기로 결정.

## 예: 다양한 class들을 하나로 통일

예: 자연어쿼리를 Task로 classify 하는 경우

[Set Alarm in 30 mins] → Alarm, Timer, Phone,



자연어처리  
라벨링에서는 언어가  
의미하는

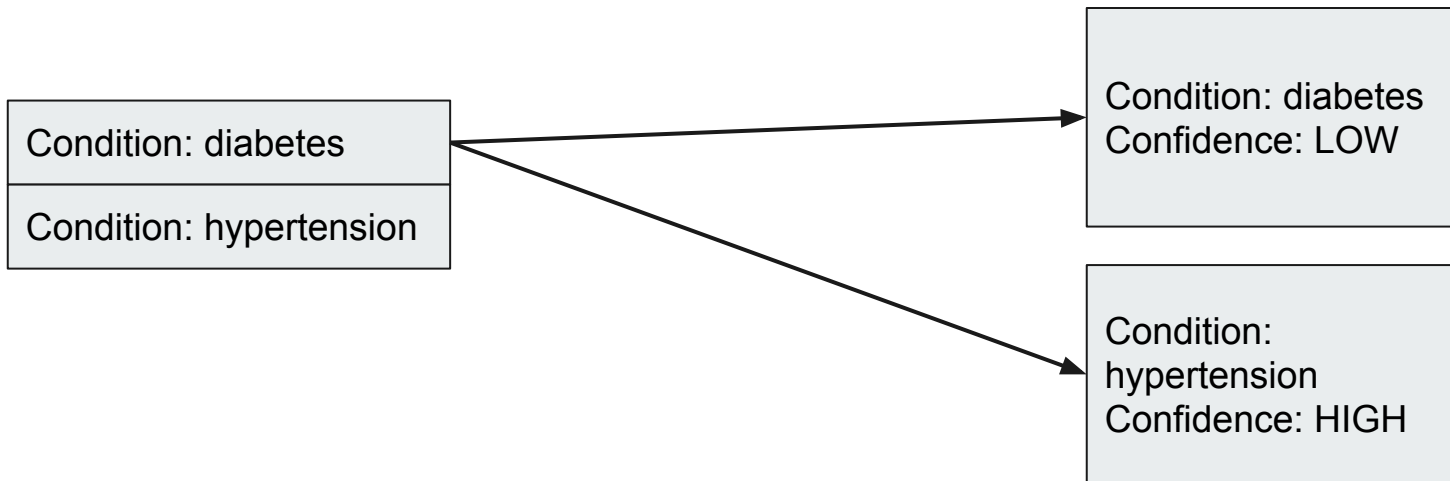
**semantic** 그대로  
데이터를 수집하고,  
차후

- post processing이 후 학습
- 그대로 학습 후 business logic

## 예: 더 많은 class들을 정의

예: 자연어를 classify 하는 경우

“ptn has a risk of diabetes due to hbp”



---

# Responsible AI: Data, Fairness

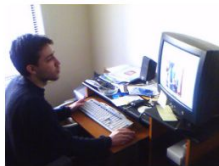
# How to collect data responsibly?



- 예) 사진들이 모여있는 데이터셋에서 “프로그래머”를 구별하는 모델을 만들고 싶다.
  - 어떤 방식으로 ground truth 데이터 수집을 진행할까? 어디서?
1. 웹에서 [프로그래머], [programmer]라고 검색. ([예1](#), [예2](#), [예3](#))
  2. 개발자 컨퍼런스, 해커쏰 (hackathon) 사진들 ([예1](#), [예2](#))
  3. 내 사진들..?

# Fair model data set: “Programmer”

Original:



**What's wrong with this picture?**

이 사진들의 공통점은?

# Fair model data set: “Programmer”

Original:



Balancing gender:





# Fair model data set

Original:



Balancing gender:



Balancing skin color:



# Fair model data set: “Programmer”

Original:



Balancing gender:



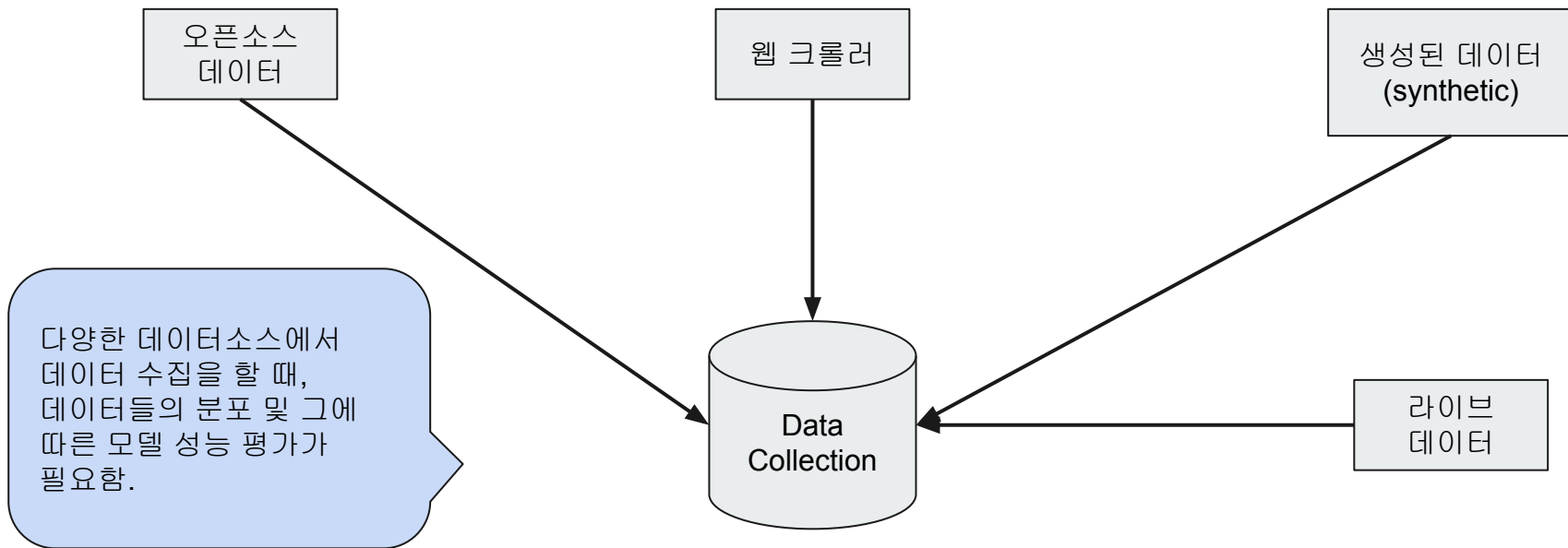
Balancing skin color:



Balancing age:



# Responsible Data Source



# ML Fairness

공정(Fair)하지 못한 머신러닝이 야기하는 문제점은?

## 1. 할당 피해(Harm by Allocation)

- ML 시스템이 특정 그룹에 대한 기회, 리소스 또는 정보를 확장하거나 알려주지 않음.

## 2. 서비스 품질 피해 (Harm by Quality of Service)

- 특정 그룹에 대해서는 다른 그룹에 대비 제품의 품질이 떨어지거나 동일하게 작동하지 않음.

예) 채용 시스템에서 서류 합격/통과 screening system에 gender bias가 있음.

→ 인터뷰까지 도달하는 지원자 풀에 allocation bias

예) 음성인식 시스템에 미국 원어민 영어 발음에 bias가 있음.

→ 비원어민에게는 음성인식이 잘 되지 않음.

# How to ensure ML Fairness?

## 학습데이터 분포 조정

- Group A, B, C 등에 해당하는 데이터셋의 크기가 서로 비슷하게 분포하여 모델링.
- 혹은 각기 따로 모델 학습

매뉴얼하게  
**protected group**을  
정의 필요

## 모델 공정도 (Fairness Indicator):

- 예: 해당 데모그래픽에 속해 있거나 그렇지 않아도 동일한 확률로 같은 **prediction**이 나옴 (Group fairness)
- Protect할 그룹을 지정해서 모니터링.

Custom protected  
**group**을 사용자가  
직접 지정 가능.

## 투명하게 모델의 정보를 공개

- Model Card: 해당 모델에 어떤 데이터셋으로 학습되었고, 어떤 데이터셋에서 성능이 저하됨을 표준화된 포맷으로 공개.
- 예: Face Detection Model Card